 DOI: https://doi.org/10.12345/bdai.v7i3.38159

Application of Data Mining Technology in Disease Risk Early Warning Based on Medical Big Data

Shengping Wan

Wuhan Wenhua University, Wuhan, Hubei, 430074, China

Abstract

The rapid accumulation of medical big data provides robust data support for disease risk early warning. However, traditional statistical methods exhibit significant limitations when processing high-dimensional, nonlinear, and multimodal data. Data mining technology, with its powerful capabilities in pattern recognition and knowledge discovery, has become a core approach for achieving early disease identification and precise prevention and control. Based on the characteristics of medical big data, this paper explores the application pathways of algorithms such as classification, clustering, association rules, and deep learning in clinical, imaging, and genomic data. Through feature extraction and model training, individualized risk prediction for chronic diseases such as cardiovascular diseases, diabetes, and malignant tumors can be achieved. The study demonstrates that the integration of data mining and artificial intelligence can significantly enhance the accuracy and intelligence level of disease early warning, providing a scientific basis for the construction of smart healthcare systems and public health management.

Keywords

medical big data; data mining; disease risk prediction; machine learning; health management

数据挖掘技术在医疗大数据疾病风险预警中的应用

万盛萍

武汉文华学院，中国·湖北 武汉 430074

摘　要

医疗大数据的迅速积累为疾病风险预警提供了坚实的数据支撑，但传统统计方法在处理高维度、非线性及多模态数据时存在明显不足。数据挖掘技术凭借强大的模式识别与知识发现能力，成为实现疾病早期识别与精准防控的核心手段。本文基于医疗大数据特征，探讨分类、聚类、关联规则与深度学习等算法在临床、影像及基因组数据中的应用路径。通过特征提取与模型训练，可实现对心血管疾病、糖尿病及恶性肿瘤等慢性疾病的个体化风险预测。研究表明，数据挖掘与人工智能的融合能显著提升疾病预警的准确性与智能化水平，为智慧医疗体系建设与公共卫生管理提供科学依据。

关键词

医疗大数据；数据挖掘；疾病风险预警；机器学习；健康管理

1 引言

随着信息化医疗与智能健康管理的发展，医院信息系统、可穿戴设备、电子病历及基因测序平台不断积累海量健康数据。医疗大数据具有体量庞大、类型复杂、更新迅速和关联性强等特征，其蕴含的隐性规律对疾病早期识别与风险预测具有重大意义。数据挖掘技术通过算法模型对大规模数据进行模式识别、异常检测与知识提取，能够在非结构化医疗信息中发现潜在的风险因素。当前，数据挖掘在疾病预警系统、智能诊疗与公共卫生管理中展现出广阔的应用前景。本文拟从数据挖掘技术体系出发，探讨其在医疗大数据环境下的应用机制与技术路径，并分析其在疾病风险预警中的典型案例与发展趋势，以期为健康信息化领域提供理论指导与实践参考。

【作者简介】万盛萍（1983–），女，中国湖北武汉人，硕士，助教，从事数据挖掘与数据仓库研究。

2 医疗大数据与疾病风险预警的理论基础

2.1 医疗大数据的特征与构成

医疗大数据具有典型的“4V”特征，即体量巨大（Volume）、类型多样（Variety）、更新迅速（Velocity）与价值密度低（Value low）。其数据来源包括电子健康档案（EHR）、医学影像、可穿戴设备监测数据、基因组测序、药物反应记录及社会行为信息等。EHR 数据记录了患者诊疗全过程，是疾病建模的核心基础；影像数据与生理信号可提供时空变化信息，支持早期病理识别；基因组与代谢组数据揭示分子水平的个体差异，为精准预测提供微观依据。由于医疗数据结构复杂、格式不统一，往往存在数据冗余、缺失与噪声问题。实现高质量的数据融合需依托数据清洗、语

义标准化与特征工程，通过自然语言处理（NLP）与多模态融合算法构建统一的数据表达模型，为疾病预测与智能诊断提供可靠的数据基础。

2.2 疾病风险预警的概念与模型框架

疾病风险预警是一种基于数据挖掘的预测性医学模式，旨在利用大规模健康数据识别高危人群并实现疾病的早期干预。其核心思想是通过算法学习患者特征与疾病发生概率之间的映射关系，建立个体化风险评估模型。典型模型包括逻辑回归、支持向量机（SVM）、决策树、随机森林以及深度神经网络（DNN）。这些模型通过输入历史病例特征数据，实现对未来疾病风险的定量预测。风险预警系统一般包含数据采集、特征提取、模型训练与结果可视化四个环节。在评估阶段，常采用ROC曲线、AUC值、灵敏度（Sensitivity）与特异度（Specificity）等指标综合判断模型性能。高性能的预警模型不仅能提升医疗服务效率，还可支持政府和医疗机构制定疾病防控策略，推动医疗体系由“被动治疗”向“主动预防”转变。

2.3 数据挖掘技术在风险预测中的优势

数据挖掘技术能够突破传统统计模型在高维数据处理与非线性关系建模中的局限，具有更强的模式识别与预测能力。在医疗风险预警中，数据挖掘通过机器学习与人工智能算法，从海量、复杂的临床数据中自动提取潜在规律。聚类分析可实现患者分层管理与疾病亚型识别，揭示个体间的生理差异；关联规则挖掘可揭示多病共现与病因耦合关系，为病理机制研究提供依据；分类算法可用于疾病风险分级与早期诊断决策。与传统统计方法不同，数据挖掘具备自学习与自适应特征，可动态更新模型参数以应对实时数据变化。结合知识图谱与专家系统，可形成“数据—知识—决策”一体化的医学智能体系，从而显著提高疾病风险预测的科学性与智能化水平，为精准医疗与公共健康管理提供重要技术支撑。

3 数据挖掘算法在疾病风险预警中的应用机制

3.1 分类算法的临床应用

分类算法在医疗数据挖掘中具有核心地位，能够通过对既往病例的特征学习，实现疾病的智能化分类与预测。决策树和随机森林模型因其结构直观、计算效率高及结果易解释等特点，被广泛应用于心血管疾病、糖尿病等慢性病风险评估。支持向量机（SVM）在处理高维、非线性数据时表现出优越的泛化能力，尤其适用于乳腺癌、肺癌等早期筛查的特征判别。近年来，深度学习模型的引入极大提升了分类精度。卷积神经网络（CNN）可自动提取医学影像特征，实现病灶区域识别与恶性程度判断；循环神经网络（RNN）在时间序列建模中具有突出优势，适用于生命体征与动态监测数据分析。通过模型集成与特征加权策略，分类算法能够实现对多病种、多维度数据的同步预测，为临床早筛与个性化诊疗提供决策支持。

3.2 聚类分析与患者群体分层

聚类算法能够在无监督条件下自动识别具有相似特征的患者群体，揭示疾病内部的异质性与亚型分布。K-means算法常用于糖尿病、高血压等代谢类疾病的数据挖掘，可将患者细分为不同病理机制亚群，如“胰岛素抵抗型”与“β 细胞功能障碍型”，从而指导精准干预。基于密度的DBSCAN算法在噪声环境下表现优越，能够识别罕见病人群与异常个体，为特殊病因研究提供数据基础。层次聚类算法则可从多尺度角度揭示疾病进展的阶段性特征。聚类结果结合临床指标分析，可用于确定关键危险因素与分层管理策略。在公共卫生层面，聚类分析还可与流行病学模型耦合，用于研究疾病传播路径与人群健康模式，为群体防控策略制定提供科学依据。

3.3 关联规则与特征关系发现

关联规则挖掘作为数据挖掘的重要分支，旨在揭示不同临床特征、检验指标与生活行为之间的潜在逻辑关系。经典算法如Apriori与FP-growth可在大规模医疗数据库中提取高置信度的规则模式，用以揭示疾病的多因共现规律。例如，通过对慢性病人群的分析，可发现肥胖、高血脂与高血压之间的耦合关系，为预防性健康干预提供量化依据。关联规则通过“支持度”“置信度”和“提升度”等指标量化特征依赖强度，实现疾病成因的可解释化建模。近年来，关联规则挖掘结果常被嵌入机器学习框架中，作为辅助特征集提升模型预测性能。该方法不仅有助于发现未知病理机制，也为医生提供临床决策支持与精准用药建议，使医疗数据从被动存储转向主动知识生成，推动医学研究进入智能化阶段。

4 医疗数据处理与模型构建技术路径

4.1 数据清洗与特征提取

医疗数据来源广泛，涵盖电子病历、检验结果、影像资料与可穿戴设备监测数据，常存在缺失值、异常值及格式不统一等问题。数据清洗是确保分析质量的首要步骤，需进行多层次的数据验证与去噪。通过异常检测算法识别离群点，采用插值法或多重填补策略处理缺失值，可显著降低模型偏差。数据归一化与标准化能够避免量纲差异导致的训练失衡。特征提取是模型性能优化的关键，可利用主成分分析（PCA）降维，或通过LASSO、XGBoost特征选择算法筛选高贡献变量。在深度学习框架中，嵌入式神经网络可自动提取潜在特征表示，实现结构化与非结构化数据（如文本与影像）的协同建模。通过多源数据融合与特征编码，模型能够更准确地捕捉复杂病理模式与个体差异，为疾病风险预测提供坚实的数据基础。

4.2 多模态数据融合建模

医疗信息具有高度异构性，单一模态难以全面反映疾病的发生机制。多模态数据融合通过整合影像、基因组及临

床数据，可在宏观与微观层面同时建模。张量融合网络（TFN）可在特征层实现跨模态关联学习，通过张量分解与注意力机制捕捉模态间的潜在依赖关系。实验表明，将 CT 影像特征与实验室指标相结合，可显著提升肺癌早期风险预测的准确性。图神经网络（GNN）在多模态信息整合中表现突出，能够建模患者、特征及疾病间的复杂图结构，实现语义层面的信息耦合。通过这种融合策略，疾病预测模型从单维分析转向系统化推理，提升了预警的准确度与泛化能力。

4.3 模型优化与性能评估

高精度的医疗预测模型需兼顾计算性能与临床可解释性。模型优化环节主要通过交叉验证与正则化方法防止过拟合，并利用网格搜索或贝叶斯优化进行参数调优。集成学习（如 Bagging 与 Boosting）可进一步增强模型稳健性。模型评估应结合医学场景设定多维指标体系，包括 ROC 曲线、AUC 值、F1 分数及 Precision-Recall 平衡，全面反映预测灵敏度与特异性。为提升临床可用性，应引入可解释人工智能（XAI）技术，对关键特征贡献度进行可视化呈现，使医生能够理解模型决策逻辑。通过 LIME、SHAP 等方法实现特征重要性排序，不仅增强了模型透明度，也提升了临床信任度。持续模型优化与动态评估机制的建立，使数据挖掘成果更好地转化为临床辅助决策工具，实现精准预警与智能诊疗的统一。

5 数据挖掘在典型疾病风险预警中的应用案例

5.1 心血管疾病风险智能预测

在心血管疾病风险预警领域，数据挖掘技术已成为精准医学的重要支撑。基于机器学习的多变量预测模型能够整合血压、血脂、血糖、体质指数及生活方式等多源数据，建立个体化风险评分体系。研究表明，采用随机森林与逻辑回归混合建模方法，其预测冠心病和心肌梗死发生概率的 AUC 值可达 0.92 以上，显著优于传统 Framingham 风险评分模型。该模型可自动识别潜在高危人群，结合特征重要性分析揭示关键危险因素，如高甘油三酯、高血压及吸烟等行为习惯的交互效应。系统化应用于社区体检数据后，可实现早期预警与个性化干预，为公共健康管理与慢病防控提供科学依据。通过实时数据反馈机制，模型还能动态更新个体风险变化趋势，实现连续监测与分级管理，显著提升心血管疾病防治的前瞻性与精确性。

5.2 糖尿病及并发症风险评估

糖尿病的慢性化进程与个体生活方式密切相关，传统静态评估方法难以捕捉动态风险变化。基于深度学习与时间序列分析的数据挖掘模型，能够融合血糖监测、饮食记录、运动强度及药物依从性数据，实现长期趋势预测。循环神经网络（RNN）与长短期记忆网络（LSTM）通过捕捉时序特征，能够有效预测血糖波动并识别糖尿病视网膜病变、肾功能损伤等并发症的高危个体。在社区医疗系统中部署后，模型可对患者异常血糖趋势发出早期警报，指导个性化用药与饮食调整。临床验证显示，该系统可使高血糖事件发生率降低约 30%，并减少了住院频率。数据挖掘技术的引入，实现了糖尿病管理从被动监测向主动预防的转变，推动了慢病管理体系的智能化与精细化发展。

5.3 恶性肿瘤早期筛查与影像识别

恶性肿瘤的早期筛查是提高生存率的关键环节，数据挖掘技术在医学影像智能识别中的应用极具潜力。基于卷积神经网络（CNN）的图像分析模型可在 CT、MRI 及病理切片图像中识别微小病灶，其特征提取能力显著优于传统人工诊断。模型通过多层卷积结构提取肿瘤形态、纹理及边缘信息，实现病灶的自动分割与分类。结合基因组学与转录组数据，可进一步构建“影像—基因”联合分析框架，用以区分良恶性肿瘤及判断分期。部分 AI 辅助筛查系统已在乳腺癌、肺癌与肝癌临床筛查中投入应用，其识别精度达 95% 以上，显著缩短诊断周期。数据挖掘的多维特征融合机制不仅提升了筛查灵敏度，也为个体化肿瘤预防与精准治疗提供了数据支撑，为早诊早治模式的构建奠定了技术基础。

6 结语

数据挖掘技术在医疗大数据疾病风险预警中的应用，正推动医学从经验驱动向数据驱动转型。通过算法创新与数据融合，医学研究能够实现疾病风险的早识别与精准干预，为临床诊疗与公共卫生政策提供决策依据。未来的发展方向在于强化模型解释性、实现跨机构数据共享与隐私保护。联邦学习与差分隐私技术的引入，将促进医疗数据在安全前提下的高效利用。数据挖掘与医学知识图谱、人工智能深度学习的融合，将推动智能预警系统迈向自动化与个性化阶段。通过科学算法、完善数据治理与伦理规范的协同发展，数据挖掘将在疾病预防与健康管理中发挥更深远的价值，为建设智慧医疗体系与“健康中国”目标提供坚实技术支撑。

参考文献

[1] 魏星,葛庆伟.人工智能技术在医疗数据挖掘中的应用现状探索[J].中国高新科技,2025,(06):52-54.

[2] 林枫.云计算技术在医疗大数据挖掘平台设计中的应用[J].电脑知识与技术,2015,11(30):3-4.

[3] 黄文扬.基于机器学习的健康体检数据慢性疾病风险预测研究[D].电子科技大学,2024.